Unravelling small world networks

Unravelling Small World Networks

Desmond J. Higham

Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK.

Abstract

New classes of random graphs have recently been shown to exhibit the *small world phenomenon*—they are clustered like regular lattices and yet have small average pathlengths like traditional random graphs. Small world behaviour has been observed in a number of real life networks, and hence these random graphs represent a useful modelling tool. In particular, Grindrod [*Range Dependent Random Graphs and their Applications to Modelling Large Small World Proteome Datasets*, Preprint maths0112, University of Bath, 2001] has proposed a class of range dependent random graphs for modelling proteome networks in bioinformatics. A property of these graphs is that, when suitably ordered, most edges in the graph are short-range, in the sense that they connect near-neighbours, and relatively few are long-range. Grindrod also looked at an inverse problem—given a graph that is known to be an instance of a range dependent random graph, but with vertices in arbitrary order, can we reorder the vertices so that the short-range/long-range connectivity structure is apparent? When the graph is viewed in terms of its adjacency matrix, this becomes a problem in sparse matrix theory: find a symmetric row/column reordering that places most nonzeros close to the diagonal. Algorithms of this general nature have been proposed for other purposes, most notably for reordering to reduce fill-in and for clustering large data sets. Here, we investigate their use in the small world reordering problem. Our numerical results suggest that a spectral reordering algorithm is extremely promising, and we give some theoretical justification for this observation via the maximum likelihood principle.

Key words: adjacency matrix, bandwidth, bioinformatics, Cuthill-McKee, envelope, genome datasets, random graph, Laplacian, maximum likelihood, minimum degree, reordering, small world phenomenon, sparse matrix, two-sum.

Email address: djh@maths.strath.ac.uk (Desmond J. Higham).

URL: http://www.maths.strath.ac.uk/ aas96106/ (Desmond J. Higham).

[1] This work was carried out while I visited The Fields Institute for Research in Mathematical Sciences, Toronto, from January to July, 2002. I acknowledge the support of a Research Fellowship from The Leverhulme Trust.

1 Introduction

We consider a graph with vertices labelled $v_1, v_2, \ldots, v_N$ and suppose that a directed edge connecting v_i to v_j exists with probability $p_{ij} := f(|j-i|)$ for some suitable function f. Equivalently, regarding the graph as being defined in terms of its adjacency matrix, we consider matrices $A \in \mathbb{R}^{N\times N}$ such that, independently for each entry, $a_{ij} = 1$ with probability $f(|j-i|)$ and $a_{ij} = 0$ otherwise. Such *range dependent* random graphs were introduced and analysed by Grindrod [9]. They can exhibit the *small world phenomenon*—high local clustering coupled with relatively short expected pathlengths—that was first studied by Watts and Strogatz [28]. Many real life networks have been found to exhibit the small world phenomenon [1,8,20,26–28], and random graphs that capture this effect form useful models for simulation and analysis [3,6,11,12,17,21,24]. Grindrod [9] developed the range dependent random graph framework as a means to model the "many to many" connections that have been observed in experiments on gene to gene and protein to protein interactions [2,5,10,14–16].

Grindrod focussed on edge probability functions with the power law form

$$f(k) = \alpha\lambda^{k-1}, \tag{1}$$

where $\alpha, \beta \in (0, 1]$ are constants. Under the natural ordering, $\{v_1, v_2, \ldots, v_N\}$, choosing $\alpha = 1$ forces each pair of neighbours to be connected, and the parameter λ controls how quickly $f(k)$ decays, that is, how rapidly the probability of an edge reduces as a function of the distance between vertices.

Grindrod also highlighted a fascinating inverse problem. Suppose we are given a graph, that is, a list of vertices in arbitrary order and a list of edges, which is known, or suspected, to be well modelled by the range dependent class. How can we reorder the vertices in such a way that the range dependent connectivity is apparent? In the genomics data set context, such a reordering is extremely valuable, as it reveals key information about functional relationships between genes (or the proteins for which they code). The corresponding matrix computation problem is:

> Given a sparse matrix, find a symmetric row/column permutation that forces as many nonzeros as possible to be close to the diagonal.

The problem can be made precise by defining an objective function $F : \mathbb{R}^{N\times N} \mapsto \mathbb{R}^+$ that measures "closeness to the diagonal of elements in a matrix". In general, minimizing $F(PAP)$ over all permutation matrices P is, of course, a very difficult problem in combinatoric optimization. Grindrod outlined a heuristic approach where F is defined via the maximum likelihood

principle. In this work we give a preliminary investigation into the use of existing algorithms that have been designed for related tasks in sparse matrix computation.

The basic problem that we are tackling is illustrated in Figure 1. The left-hand picture shows the nonzeros in an instance of the random graph defined by (1) with $N = 200$, $\lambda = 0.9$ and $\alpha = 1$. Note that most nonzeros are clustered towards the diagonal, but a few 'long-range' nonzeros have been produced. In the right-hand picture we have randomly reordered the vertices; that is, we show the nonzero pattern in a matrix PAP; where P is an arbitrary permutation matrix. Given the picture on the right, our task is to find the reordering that produces something close to the picture on the left.

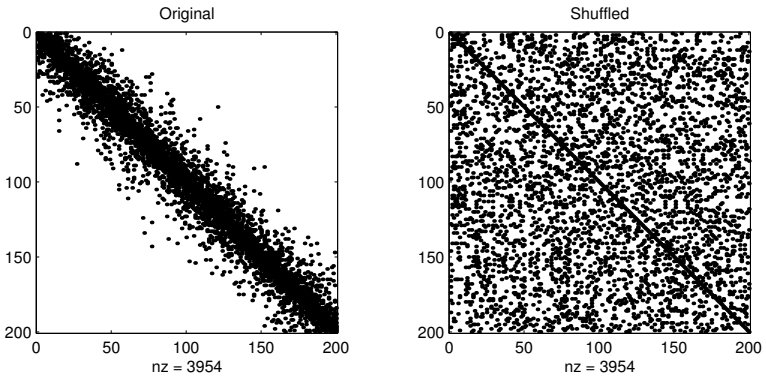

Fig. 1. Partially random graph (1) with $N = 200$, $\lambda = .9$ and $\alpha = 1$. Left original, Right with symmetric row and column shuffling.

In the next section we introduce the reordering algorithms that are to be tested. §3 gives numerical results. We interpret the results and draw some conclusions in §4.

2 Algorithms

2.1 Symmetric Reverse Cuthill-McKee and Symmetric Minimum Degree

Reordering to avoid fill-in during subsequent factorizations is a classic issue in sparse matrix computation. The two algorithms that we consider here, *Symmetric Reverse Cuthill-McKee* and *Symmetric Minimum Degree* are particularly popular tools. Symmetric Reverse Cuthill-McKee is especially promising in our context, as it may be regarded as a heuristic attempt to find a reordering that minimizes the *bandwidth*, $\max\{|i-j| : a_{ij} \neq 0\}$. Since these algorithms are standard, we do not describe them here. Details can be found, for example in [22].

An alternative reordering algorithm for sparse matrices was proposed in [4]. We note that these ideas have been further pursued in [18,23,25]. To describe the algorithm, we assume that the matrix A is symmetric. The task considered in [4] is to reduce the *envelope*, which is defined as the number of nonzeros, plus the number of zeros on each row that lie between nonzeros; that is $\sum_{ij} f_{ij}$, where $F \in \mathbb{R}^{N\times N}$ is defined by $f_{ij} = 1$ if $a_{ij} \neq 0$ or if there exist j_1, j_2 such that $j_1 \leq j \leq j_2$ with $a_{i,j_1} a_{i,j_2} \neq 0$, and $f_{ij} = 0$ otherwise. However, the algorithm is motivated in [4] as an attempt to minimize the *two-sum*

$$\sum_{\{i,j:a_{ij}\neq 0\}} (i-j)^2.$$

We reproduce the argument below, as it is helpful when interpreting the numerical results. Further justification for the algorithm appears in [7].

Using $\mathcal{P}$ to denote the set of permutations of the integers $\{1, 2, 3, \ldots, N\}$, reordering to minimize the two-sum means solving

$$\min_{x\in\mathcal{P}} \left\{ \sum_{\{i,j:a_{ij}\neq 0\}} (x_i - x_j)^2 \right\}. \tag{2}$$

This can be reduced to the problem

$$\min_{x\in\mathcal{P}} x^T Q x, \tag{3}$$

where the *Laplacian matrix* Q is defined by

$$q_{ij} = \begin{cases} -1 & \text{for } i \neq j \text{ and } a_{ij} \neq 0, \\ 0 & \text{for } i \neq j \text{ and } a_{ij} = 0, \\ -\sum_{j=1, j\neq i}^{n} q_{ij} & \text{for } i = j. \end{cases}$$

Now a heuristic is introduced that makes the problem tractable, at the expense of computing a guaranteed optimal solution. Instead of minimizing over the discrete set $\mathcal{P}$, relax the problem (3) to $x \in \mathbb{R}^N$ and factor out the trivial solutions $x = \mathbf{0}$ and $x = \mathbf{e}$, where $\mathbf{0} = [0, 0, \ldots, 0]^T$ and $\mathbf{e} = [1, 1, \ldots, 1]^T$. This leads us to

$$\min_{\{x\in\mathbb{R}^N : x^T\mathbf{e}=0, \|x\|_2=1\}} x^T Q x, \tag{4}$$

which is solved by taking x to be the eigenvector $x^{[2]}$ corresponding to the second smallest eigenvalue of Q. Although the "solution", $x^{[2]}$, is a real-valued vector rather than a permutation vector, we can use the ordering of the elements in $x^{[2]}$ to induce a permutation vector $p \in \mathcal{P}$. So we choose a $p \in \mathcal{P}$ such that $p_i \leq p_j$ if and only if $x_i^{[2]} \leq x_j^{[2]}$. Applying this reordering to A is what we mean by the *Spectral Reordering* algorithm.

3 Numerical Experiments

Now we give some computational results. These were generated with MATLAB (Version 6.0.0.88 (R12)) [13,19]. We used MATLAB's built-in implementations of Symmetric Reverse Cuthill-McKee and Symmetric Minimum Degree, which are provided through `symrcm.m` and `symmmd.m`, respectively. Typing `help` on these two functions produces the following descriptions:

```
>> help symrcm

 SYMRCM Symmetric reverse Cuthill-McKee permutation.
    p = SYMRCM(S) returns a permutation vector p such that S(p,p)
    tends to have its diagonal elements closer to the diagonal than S.
    This is a good preordering for LU or Cholesky factorization of
    matrices that come from "long, skinny" problems.  It works for
    both symmetric and asymmetric S.

    See also SYMMMD, COLMMD, COLPERM.

>> help symmmd

 SYMMMD Symmetric minimum degree permutation.
    p = SYMMMD(S), for a symmetric positive definite matrix S,
    returns the permutation vector p such that S(p,p) tends to have a
    sparser Cholesky factor than S.  Sometimes SYMMMD works well
    for symmetric indefinite matrices too.

    See also COLMMD, COLPERM, SYMRCM.
```

Each experiment is presented as eight pictures. The top-left picture gives an adjacency matrix computed as an instance of a range dependent random graph. The top-right picture shows a shuffled version of the matrix. This is the data matrix to which the algorithms are applied. The left-hand plot in the second row shows the data matrix reordered according to the Symmetric Reverse Cuthill-McKee algorithm. The left-hand plot in the third row compares the ordering produced by this algorithm with the "correct" ordering;

that is the ordering that recovers the original matrix. More precisely, we plot $p(q_1), p(q_2), \ldots, p(q_N)$, where $p \in \mathcal{P}$ represents the original shuffle and $q \in \mathcal{P}$ is the permutation from Symmetric Reverse Cuthill-McKee. In MATLAB this is `plot(p(q))`. For this picture a straight line of slope $+1$ indicates a perfect reconstruction of the original matrix. Because $\{N, N-1, N-2, \ldots, 1\}$ is as good as $\{N, N-1, N-2, \ldots, 1\}$ in terms of identifying neighbouring vertices, a line of slope -1 is equally acceptable. Deviations from a straight line indicate a mismatch between the original shuffling and the unshuffling that was reverse engineered by the algorithm. Similarly, the remaining pictures in rows two and three give the same information for Symmetric Minimum Degree and Spectral Reordering, respectively. Because Spectral Reordering is designed exclusively for symmetric matrices, we applied the algorithm to the matrix B, where $b_{ij} = 1$ if and only if either $a_{ij} = 1$ or $a_{ji} = 1$, where A is the shuffled data mtarix. To get the right-hand picture in row two, the computed ordering was applied to A.

In all tests, we used the power law decay form (1) with $\alpha = 1$ and took $N = 600$ vertices. (Formally, we also redefined $f(0) = 1$, so that probabilities do not exceed 1.)

Figures 2, 3 and 4 correspond to the cases $\lambda = 0.8$, $\lambda = 0.9$ and $\lambda = 0.975$, respectively.

It may be argued that typical protein-protein or gene-gene interaction networks correspond to undirected graphs, and hence to symmetric adjacency matrices. To test the algorithms on symmetric versions of range dependent random graphs, we computed matrices according to the rule $a_{ij} = 1$ with probability $f(|j - i|)$ and $a_{ij} = 0$ otherwise for $j \leq i$, and $a_{ij} = a_{ji}$ for $j > i$. Corresponding results for the parameter values used above appear in Figures 5–7.

For each experiment, we also computed the bandwidth, envelope and two-sum of the original matrix and the three matrices resulting from the algorithms. Tables 1— 3 give the results.

4 Conclusions and Observations

We draw the following points from the numerical results of the previous section.

(1) The algorithms behave similarly on symmetric and unsymmetric problems.
(2) The Symmetric Minimum Degree algorithm is not successful at recover-

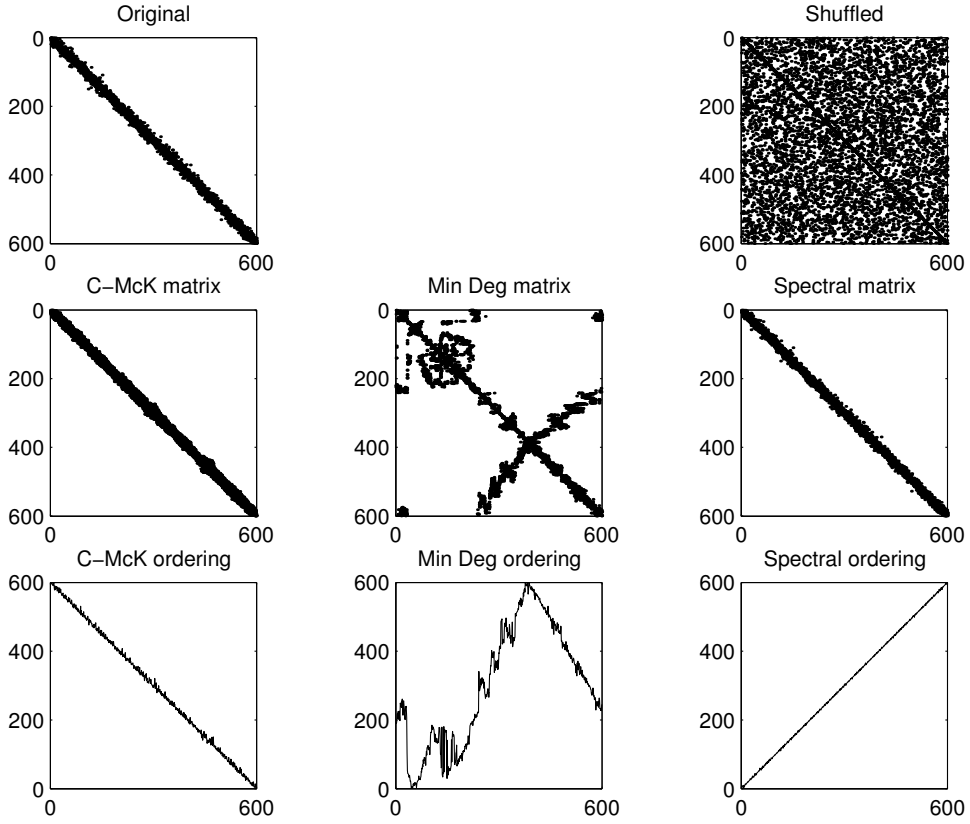

Fig. 2. Unsymmetric, $\lambda = 0.8$. See text for details.

	Original Matrix	Symm. Rev. C-McK	Symm. Min. Deg.	Spectral
$\lambda = 0.8$, Unsymm.	42	31	591	41
$\lambda = 0.8$, Symm.	40	24	599	31
$\lambda = 0.9$, Unsymm.	110	65	599	96
$\lambda = 0.9$, Symm.	78	64	597	69
$\lambda = 0.975$, Unsymm.	455	284	597	442
$\lambda = 0.975$, Symm.	387	285	598	360

Table 1
Bandwidth

ing the original range dependent connectivity structure in the adjacency matrix. However, it does seem to pick up some ordering information.

(3) Symmetric Reverse Cuthill-McKee is fairly successful at reproducing the original data matrix for the smaller λ values of 0.8 and 0.9; that is, for matrices where there are relatively few long-range connections. It is less successful for the weaker decay rate of $\lambda = 0.975$, although even in this case a lot of information is carried through.

(4) Spectral Reordering is the most promising of the three algorithms, and makes a very accurate job of undoing the initial shuffling.

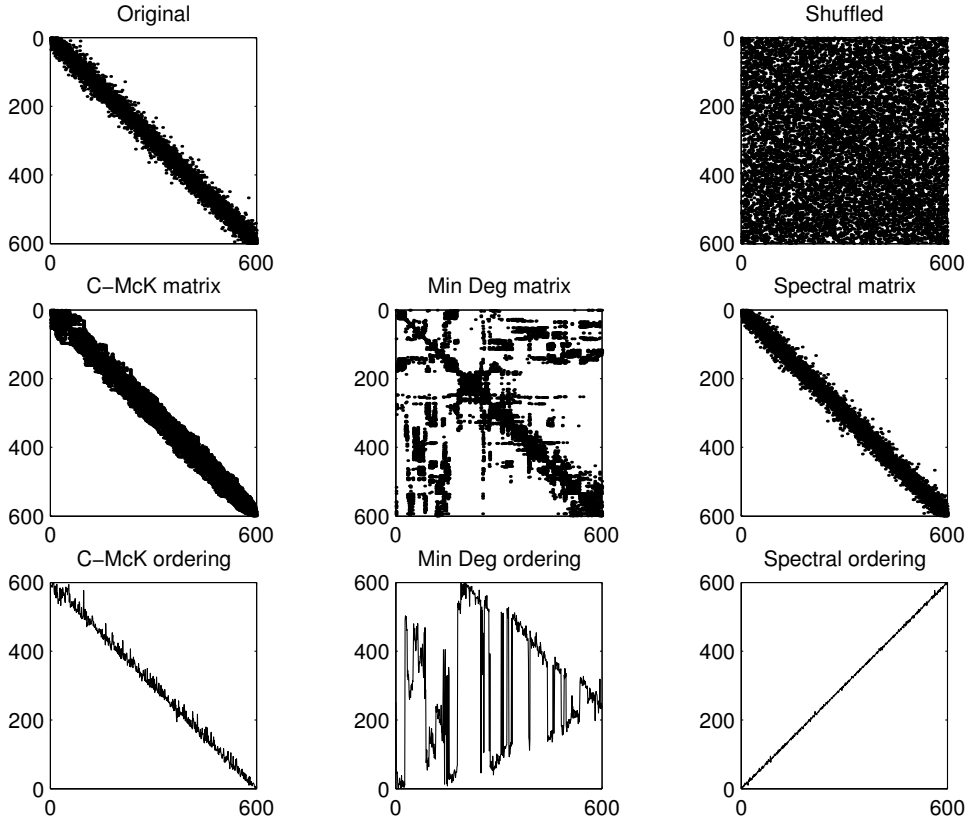

Fig. 3. Unsymmetric, $\lambda = 0.9$. See text for details.

	Original Matrix	Symm. Rev. C-McK	Symm. Min. Deg.	Spectral
$\lambda = 0.8$, Unsymm.	1.3	1.6	11.7	1.3
$\lambda = 0.8$, Symm.	1.3	1.6	10.7	1.3
$\lambda = 0.9$, Unsymm.	3.3	4.1	21.4	3.3
$\lambda = 0.9$, Symm.	3.4	4.2	12.6	3.3
$\lambda = 0.975$, Unsymm.	16.8	19.0	33.2	16.6
$\lambda = 0.975$, Symm.	16.6	19.7	30.5	16.4

Table 2
Envelope, scaled by 10^4 and rounded to 1 decimal place.

(5) It is clear from the pictures, and from Table 2, that Symmetric Reverse Cuthill-McKee tends to focus on reducing the envelope at the expense of generally shepherding all elements towards the diagonal. In our context the original matrix may well have "outliers" that represent genuine long-range contacts and hence should be left as such. Since the envelope is not tolerant to outliers, it is not the most suitable basis for a reordering.

(6) The two-sum appears to be quite a robust objective function for the range of λ values used here, and the relaxation from permutations in (2)

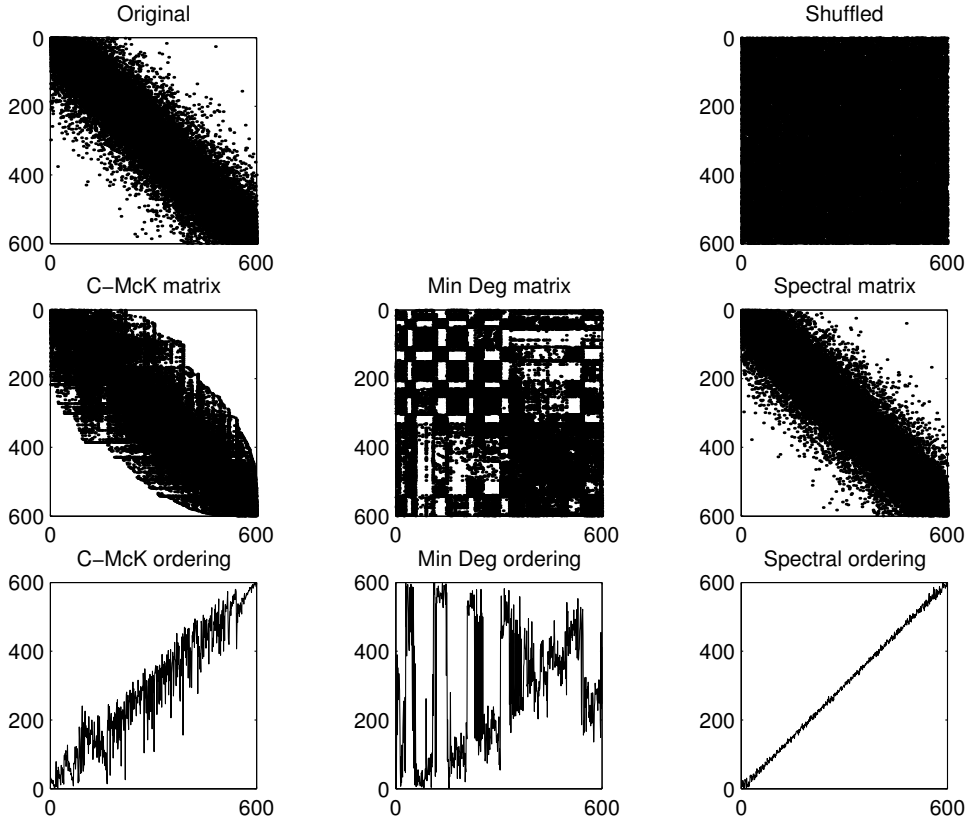

Fig. 4. Unsymmetric, $\lambda = 0.975$. See text for details.

	Original Matrix	Symm. Rev. C-McK	Symm. Min. Deg.	Spectral
$\lambda = 0.8$, Unsymm.	2.7	5.3	1588.4	2.6
$\lambda = 0.8$, Symm.	2.7	5.1	1298.9	2.4
$\lambda = 0.9$, Unsymm.	21.4	52.8	5339.7	20.9
$\lambda = 0.9$, Symm.	22.5	56.3	2105.7	21.1
$\lambda = 0.975$, Unsymm.	1220.6	2974.3	17025.6	1212.1
$\lambda = 0.975$, Symm.	1210.5	3629.8	15922.0	1186.2

Table 3
Two-sum, scaled by 10^5 and rounded to 1 decimal place.

to real vectors in (4) still leads to good solutions. In several cases, Spectral Reordering pushes the two-sum below its value for the the original matrix, but produces a very similar ordering.

We point out that these preliminary results are based on single instances of range dependent graphs. More authoritative conclusions can only be drawn from a statistical analysis based on many samples.

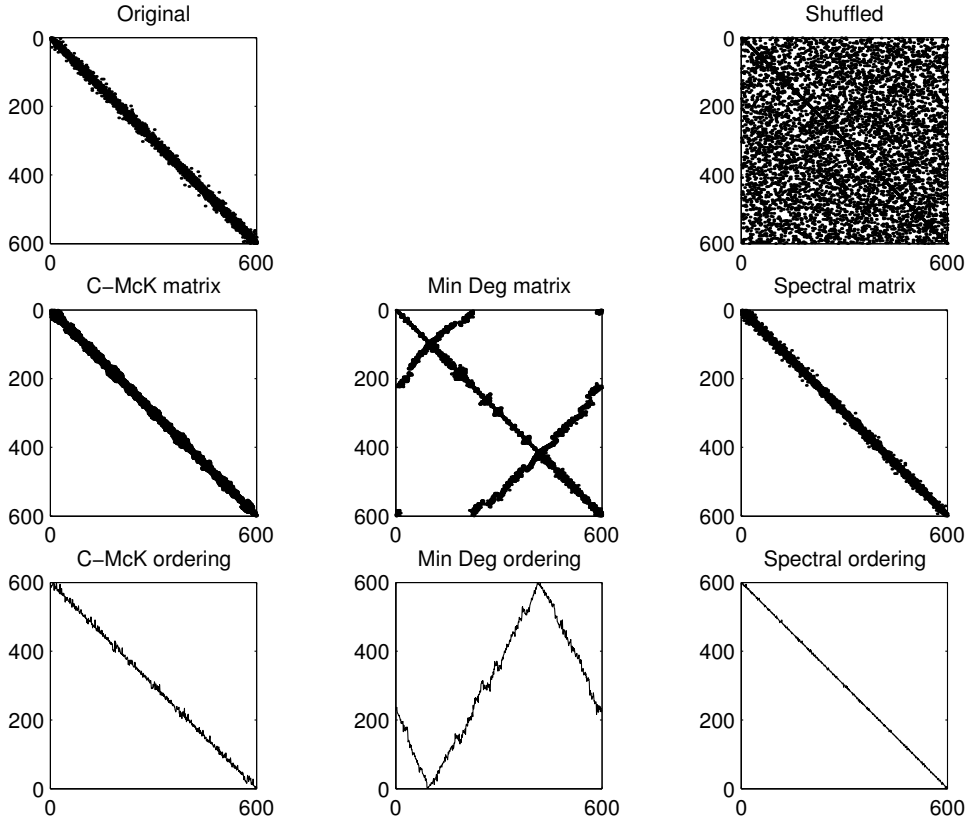

Fig. 5. Symmetric, $\lambda = 0.8$. See text for details.

As mentioned in §1, Grindrod [9] suggested using a maximum likelihood approach in order to obtain an objective function that can be minimized. Given that the correctly reordered data matrix comes from a range dependent random graph generated by a function f, the *maximum likelihood* ordering, that is, the ordering producing a matrix that has the highest probability of arising, is given by solving

$$\max_{x\in\mathcal{P}}\left\{\prod_{\{i,j:a_{ij}\neq 0\}} f(|x_i - x_j|) \times \prod_{\{i,j:a_{ij}=0\}} (1 - f(|x_i - x_j|))\right\}.$$

Grindrod noticed that this problem can be re-written

$$\max_{x\in\mathcal{P}}\left\{\prod_{\{i,j:a_{ij}\neq 0\}} \frac{f(|x_i - x_j|)}{1 - f(|x_i - x_j|)} \times \prod_{i,j} (1 - f(|x_i - x_j|))\right\}.$$

The second product inside the braces is the probability of a null graph, that is, a graph with no edges, and this is constant for all $x \in \mathcal{P}$. Hence, the maximum

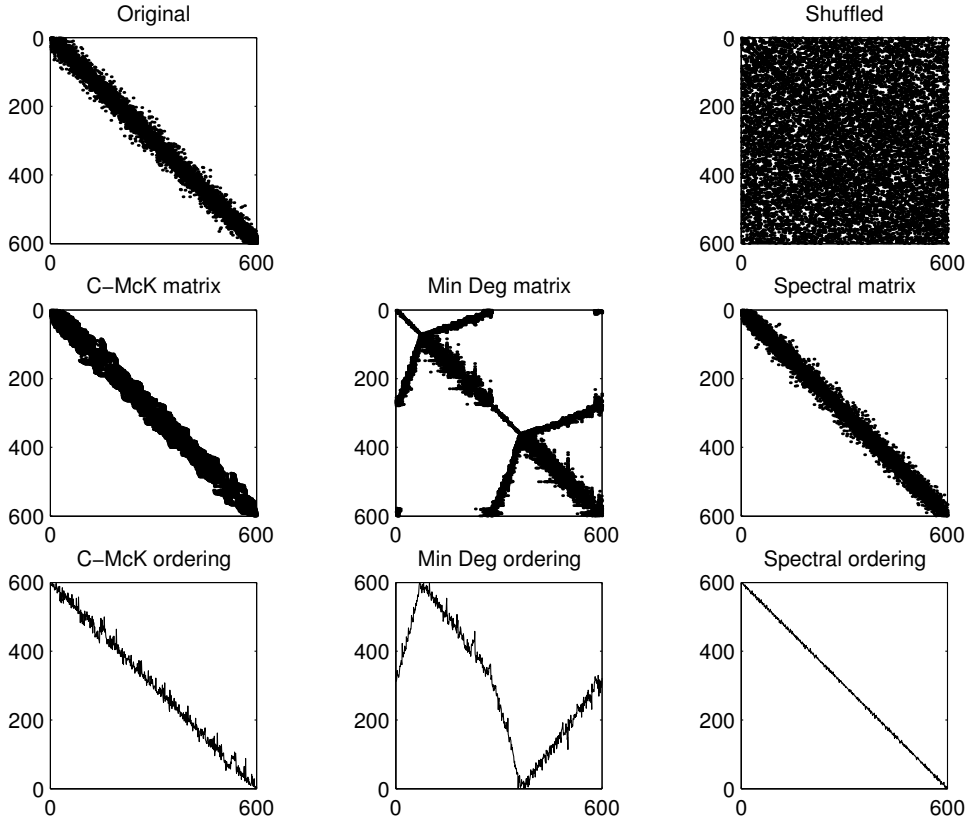

Fig. 6. Symmetric, $\lambda = 0.9$. See text for details.

likelihood ordering can be found by solving

$$\max_{x \in \mathcal{P}} \left\{ \prod_{\{i,j:a_{ij} \neq 0\}} \frac{f(|x_i - x_j|)}{1 - f(|x_i - x_j|)} \right\}. \tag{5}$$

(In practice, any edges that exist with probability 1, that is, where $f(|x_i - x_j|) = 1$, would be treated specially.) This approach has the benefit of allowing the objective function to be tuned to the data. For example, using the class (1), values for the parameters α and λ could be estimated from A. However, (5) is a hard combinatorial optimization problem, in general. Grindrod [9] outlined a hierarchical algorithm, based on iteratively improving a current guess for the best x, that can be used to tackle the problem directly. An advantage of the two-sum objective function is that, after relaxation to $\mathbb{R}^N$, it reduces to tractable numerical linear algebra. Of course, the Spectral Reordering solution could be fed in as an initial guess to Grindrod's direct method. In fact, the two approaches, two-sum minimization and maximum likelihood, are not unrelated, and understanding the connection between them gives a useful insight into the behaviour of Spectral Reordering, as we now show.

Lemma 1 *The problem of minimizing the two-sum (2) is equivalent to max-*

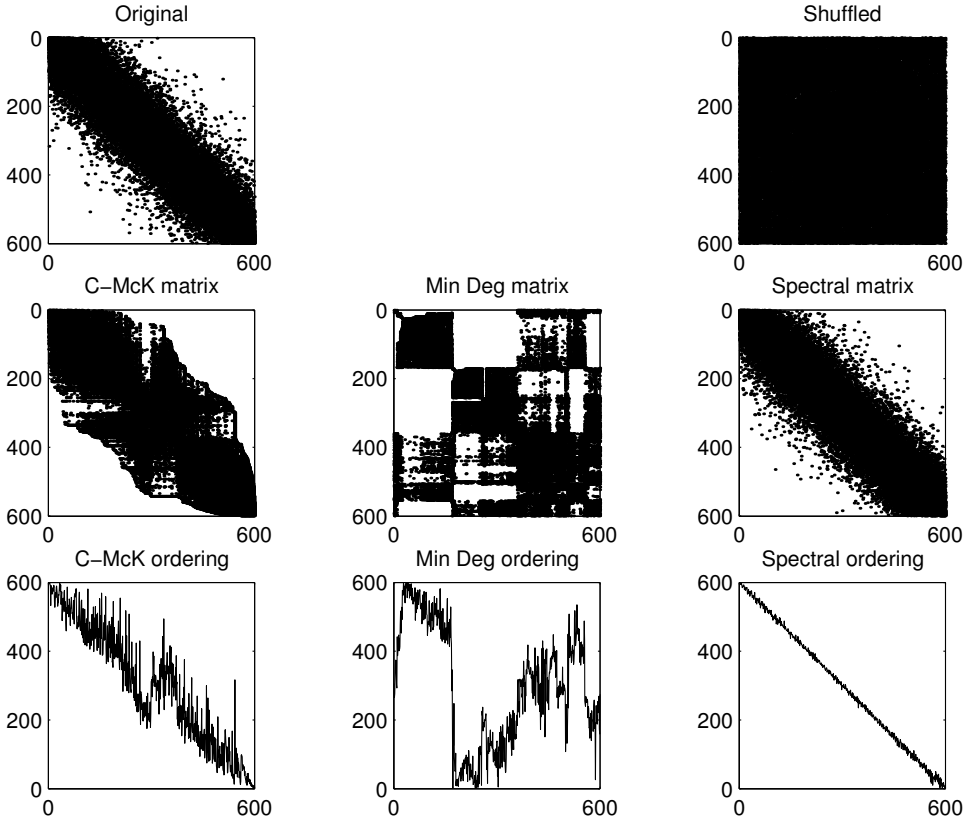

Fig. 7. Symmetric, $\lambda = 0.975$. See text for details.

imum likelihood optimization (5) with

$$f(k) = \frac{e^{-k^2}}{1 + e^{-k^2}}, \quad k = 0, 1, 2, \ldots \tag{6}$$

Proof. Multiplying by -1 in (5), the max becomes a min. Taking logs and equating the objective function with that in (2) gives the result. □

Lemma 1 shows that, from a maximum likelihood viewpoint, Spectral Reordering postulates an underlying range dependency given by f in (6). As k increases, this $f(k)$ decays faster than the geometric rate in (1). Figure 8 shows how $f(k)$ in (6) for $k \geq 1$ compares with (1) using $\alpha = 1$ and $\lambda = 0.8, 0.9, 0.975$, as in our experiments. The very rapid fall-off in (6) suggests that the Spectral Reordering approach may be less successful on data where either (a) long-range connections are not so rare or (b) long-range outliers are present due to experimental noise.

In summary, we hope that this work draws more attention to a challenging inverse problem that appears to have direct relevance to an extremely important and timely application area. There are many ways in which the ideas here (which themselves draw heavily on [9]) could be pursued. In particular, three

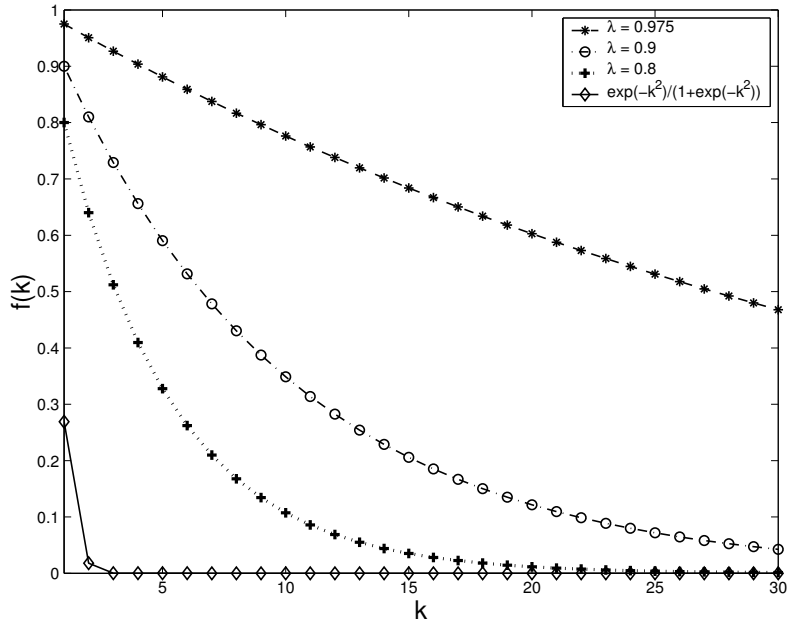

Fig. 8. Plots of $f(k)$ in (1) with $\alpha = 1$, $\lambda = 0.8, 0.9, 0.975$ and in (6).

key topics for reordering algorithms are

- large scale statistical testing on range dependent random graphs,
- experiments on large scale genome datasets,
- development of customized algorithms that combine ideas from graph theory, optimization, sparse matrix theory and statistics.

References

[1] L. Adamic, *The small world web*, Proceedings of the European Conference on Digital Libraries, (1999), pp. 443–452.

[2] R. B. Altman and S. Raychaudhuri, *Whole-genome expression analysis: challenges beyond clustering*, Current Opinion in Structural Biology, 11 (2001), pp. 340–347.

[3] A. D. Barbour and G. Reinert, *Small worlds*, Random Structures and Algorithms, 19 (2001), pp. 54–74.

[4] S. T. Barnard, A. Pothen, and H. D. Simon, *A spectral algorithm for envelope reduction of sparse matrices*, Numerical Linear Algebra with Applications, 2 (1995), pp. 317–334.

[5] M. B. Eisen, P. T. Spellman, P.O.Brown, and D.Botstein, *Cluster analysis and display of genome-wide expression patterns*, Genetics, 95 (1998), pp. 14863–14868.

[6] H. Fuks and A. T. Lawniczak, *Performance of data networks with random links*, Mathematics and Computers in Simulation, 51 (1999), pp. 103–119.

[7] A. George and A. Pothen, *An analysis of spectral envelope-reduction via quadratic assignment problems*, SIAM Journal of Matrix Analysis and its Applications, 18 (1997), pp. 706–732.

[8] P. M. Gleiss, P. F. Stadler, A. Wagner, and D. A. Fell, *Relevant cycles in chemical reaction networks*, Advances in Complex Systems, 4 (2001), pp. 207–226.

[9] P. Grindrod, *Range dependent random graphs and their applications to modelling large small world proteome datasets*, Preprint maths0112, University of Bath, Bath, 2001.

[10] T. R. Hazbun and S. Fields, *Networking proteins in yeast*, PNAS, 98 (2001), pp. 4277–4278.

[11] D. J. Higham, *A small world phenomenon for Markov chains*, Mathematics Research Report 22, University of Strathclyde, 2001.

[12] ——, *Greedy pathlengths and small world graphs*, Mathematics Research Report 8, University of Strathclyde, 2002.

[13] D. J. Higham and N. J. Higham, *MATLAB Guide*, SIAM, Philadelphia, 2000.

[14] T. Ito, T. Ichiba, O. R, M. Yoshida, M. Hattori, and Y. Sakaki, *A comprehensive two-hybrid analysis to explore the yeast protein interactome*, PNAS, 98 (2001), pp. 4569–4574.

[15] H. Jeong, S. P. Mason, A.-L. Barábasi, and Z. N. Oltval, *Lethality and centrality in protein networks*, Nature, 411 (2001), pp. 41–42.

[16] H. Jeong, B. Tombor, R. Albert, Z. Oltval, and A.-L. Barabasi, *The large scale organisation of metabolic networks*, Nature, 407 (2000), pp. 651–654.

[17] J. Kleinberg, *Navigation in a small world*, Nature, 406 (2000), p. 845.

[18] G. Kumfert and A. Pothen, *Two improved algorithms for envelope and wavefront reduction*, BIT, 35 (1997), pp. 559–590.

[19] *Using MATLAB*, The MathWorks, Inc., Natick, MA, USA. Online version.

[20] M. E. J. Newman, *The structure of scientific collaboration networks*, Proc. Natl. Acad. Sci., 98 (2001), pp. 404–409.

[21] M. E. J. Newman, C. Moore, and D. J. Watts, *Mean-field solution of the small-world network model*, Physical Review Letters, 84 (2000), pp. 3201–3204.

[22] J. Reid, I. Duff, and A. Erisman, *Direct Methods for Sparse Matrices*, Oxford University Press, 1986.

[23] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (2000), pp. 888–905.

[24] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268–276.

[25] R. Van Driessche and D. Roose, *An improved spectral bisection algorithm and its application to dynamic load balancing*, Parallel Computing, 21 (1995), pp. 29–48.

[26] A. Wagner and D. A. Fell, *The small world inside large metabolic networks*, Proc. Roy. Soc. London, B., 268 (2001), pp. 1803–1810.

[27] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, 1999.

[28] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440–442.